

Get started

Overview

You'll have an understanding of building a Docker image, deploying a Serverless endpoint, and sending a request. You'll also have a basic understanding of how to customize the handler for your use case.

Prerequisites

This section presumes you have an understanding of the terminal and can execute commands from your terminal.

RunPod

To continue with this quick start, you'll need the following from RunPod:

- RunPod account
- RunPod API Key

Docker

To build your Docker image, you'll need the following:

- Docker installed
- Docker account

GitHub

To clone the `worker-template` repo, you'll need access to the following:

- Git installed
- Permissions to clone GitHub repos

Build and push your Docker image

This step will walk you through building and pushing your Docker image to your container registry. This is useful to building custom images for your use case.

1. Clone the worker-template:

```
gh repo clone runpod-workers/worker-template
```

2. Navigate to the root of the cloned repo:

```
cd worker-template
```

3. Build the Docker image:

```
docker build --tag <username>/<repo>:<tag> .
```

4. Push your container registry:

```
docker push <username>/<repo>:<tag>
```

Now that you've pushed your container registry, you're ready to deploy your Serverless Endpoint to RunPod.


Deploy a serverless endpoint

This step will walk you through deploying a Serverless Endpoint to RunPod. You can refer to this walkthrough to deploy your own custom Docker image.

1. Login to the RunPod Serverless console.
2. Select **+ New Endpoint**.
3. Provide the following:
 - i. Endpoint name.
 - ii. Select a GPU.
 - iii. Configure the number of workers.
 - iv. (optional) Select **FlashBoot**.
 - v. (optional) Select a template.
 - vi. Enter the name of your Docker image.
 - For example `<username>/<repo>:<tag>`.
 - vii. Specify enough memory for your Docker image.

4. Select **Deploy**.

Now, let's send a request to your Endpoint.

 [Edit this page](#)